Design and Analysis of Multilevel Analytic Studies with Applications to a Study of Air Pollution

William Navidi, Duncan Thomas, Daniel Stram, and John Peters

Department of Preventive Medicine, University of Southern California, Los Angeles, California

We discuss a hybrid epidemiologic design that aims to combine two approaches to studying exposure-disease associations. The analytic approach is based on comparisons between individuals, e.g., case-control and cohort studies, and the ecologic approach is based on comparisons between groups. The analytic approach generally provides a stronger basis for inference, in part because of freedom from between-group confounding and better quality data, but the ecologic approach is less susceptible to attenuation bias from measurement error and may provide greater variability in exposure. The design we propose entails selection of a number of groups and enrollment of individuals within each group. Exposures, outcomes, confounders, and modifiers would be assessed on each individual; but additional exposure data might be available on the groups. The analysis would then combine the individual-level and the group-level comparisons, with appropriate adjustments for exposure measurement errors, and would test for compatibility between the two levels of analysis, e.g., to determine whether the associations at the individual level can account for the differences in disease rates between groups. Trade-offs between numbers of groups, numbers of individuals, and the extent of the individual and group measurement protocols are discussed in terms of design efficiency. These issues are illustrated in the context of an on-going study of the health effects of air pollution in southern California, in which 12 communities with different levels and types of pollution have been selected and 3500 school children are being enrolled in a ten-year cohort study. Exposure is being assessed through a combination of ambient monitoring, microenvironmental sampling, personal monitoring, and questionnaire data on time-activity and household characteristics. These data will be used to develop a model for personal exposures for use in the individual-level analyses, as well as for the group mean exposures for the group-level analyses. — Environ Health Perspect 102(Suppl 8):25–32 (1994)

Key words: ecologic inference, regression models, measurement error, air pollution

Introduction

Epidemiologists recognize two basic strategies for looking at the association between an exposure and a disease: ecologic studies, in which disease rates in groups of individuals are related to the average exposure rates in these groups, and analytic studies, in which individuals' disease outcomes are related to their own exposure values. Cohort studies and case-control studies are examples of the latter type. The epidemiologic literature is full of examples of discrepancies between the conclusions of the two types of studies. In a classic example, Durkheim found suicide rates in provinces of western Europe to be highly correlated with the proportion of Protestants. Regression analyses of these rates produced an estimate of the rate ratio for Protestants relative to Catholics of 7.5, compared with a value of 2 estimated on an individual basis (*1*). Similarly, numerous associations between cancer rates and mean consumption of various dietary factors have been found in ecologic correlation studies, but establishing such associations at an individual level has proven more elusive (*2*).

The resolution of such paradoxes usually turns on three issues: between-group confounding, measurement error, and restricted variability. Between-group confounding refers to a characteristic of groups that is not accounted for in the model but is the real risk factor. In the suicide example, such a factor might be the alienation felt by Catholics in predominantly Protestant provinces. This is the essential explanation of the "ecologic fallacy," in which spurious ecologic associations may be caused by a tendency for the individuals in the higher exposure groups who get the disease not to have been exposed themselves but rather to have gotten the disease as a result of some other group characteristic. Exposure measurement error has different effects on the two types of studies, generally biasing associations at the individual level toward the null, but not at the aggregate level. Finally, studies conducted within a single group may have a restricted range of variation in exposure, and hence limited power. Thus, in the diet example the positive associations at the ecologic level might be explained by some confounding variable, such as race, that is not accounted for in the analysis, whereas the lack of association at the individual level might be due to dilution of a real effect by measurement error or by restricted variability in diet within racial groups.

Each of these designs has advantages and disadvantages. The main advantage of the ecologic design is cost, but its relative freedom from measurement error bias and greater variation in exposure between groups are other advantages. On the other hand, it typically suffers from between-group confounding (partly because groups will be more heterogeneous with respect to confounders than members of groups and partly because data on confounders are unavailable) and the exposure data are usually of poor quality (e.g., food disappearance rates rather than mean intake rates). Analytic studies are more readily controlled for confounding factors and have better quality data, but may suffer from the effects of measurement error and restricted variability.

To overcome these problems, we consider a hybrid design involving aspects of both approaches, which we shall call the

This paper was presented at the 4th Japan-US Biostatistics Conference on the Study of Human Cancer held 9–11 November 1992 in Tokyo, Japan.

This research was partially supported by California Air Resources Board Contract A033-186 and Health Effects Institute Contract 8836-1486.

Address correspondence to William Navidi, Department of Preventive Medicine, University of Southern California, 1420 San Pablo Street, Los Angeles, CA 90033-9987. Telephone (213) 342-1209. Fax (213) 342-3211.

"multilevel analytic design." Key to this design is an analysis that will exploit both levels of comparison. Exposure and confounder data will be assembled on individuals, to provide the best quality possible. Individual level analyses within groups will be adjusted for measurement error. The resulting exposure-response relations then can be tested for compatibility with the between-group differences in rates; and if compatible, the two analyses can be pooled for greater power. In particular, this allows one to assess how much of the differences in disease rates between groups can be explained by differences in the distribution of risk factors.

In the next section, we provide some details about the basic design and its analysis. In the following section, we describe how the effects of measurement error may be incorporated. We then address the issue of design optimization, and provide an example with a simulation study. Finally, we describe an application to the design of the University of Sothern California (USC) study of the health effects of air pollution.

Multilevel Analytic Design and Its Analysis

The new design begins with a selection of a number of groups $g=1,...,G$, which might be defined by geographic areas (as in a study of air pollution), ethnicity (as in a study of diet), or any other factor for which group identifying data are readily available. Within each group, individuals $i=1,...,I_g$ are selected. (For notational simplicity, we set $I_g \equiv I$). Data on outcomes y_{gi}, exposures x_{gi}, and confounders v_{gi} are collected on each individual; in addition, certain characteristics of the group X_g may also be collected. For example, in an air pollution study, individual exposure information might comprise personal exposure estimates (e.g., ozone badges), microenvironment sampling (e.g., in homes, schools, cars, outdoors), or individual exposure modifying factors such as proportion of time spent outdoors or characteristics of the subjects' homes (air conditioning, presence of a smoker, heating and cooking sources, etc.). Group exposure characteristics might include estimates of the ambient levels from area monitoring. The specifics of the outcomes (continuous or binary, cross-sectional or longitudinal) and the sampling plan for individuals (survey, cohort, or case-control) will vary from study to study, but are not germane to the issues discussed here.

For conceptual and notational simplicity, we will assume that the outcome, exposure, and confounder are all univariate and continuous, and that the individuals in each group are chosen by simple random sampling. We also assume that the quantities of interest are linearly related, that is,

$$y_{gi} = \alpha_g + \beta x_{gi} + \gamma v_{gi} + \varepsilon_{gi} \qquad [1]$$

where α_g is the baseline outcome for group g and the ε_{gi} are independent random variables with $E(\varepsilon_{gi}) = 0$, $\mathrm{Var}(\varepsilon_{gi}) = \sigma^2$. Interest centers on the estimation of β, the exposure effect.

The baseline effects α_g may be considered fixed or random. Considering them random may be appropriate when the groups on which data are collected are randomly chosen from a larger population of groups. The true exposures x_{gi} and the confounders v_{gi} may also be considered either fixed or random. If the groups are randomly chosen or the subjects are randomly chosen within groups, it may be appropriate to consider them random. In what follows we will consider α_g, x_{gi}, and v_{gi} to be random, and we make the following assumptions: First, the random variables $\alpha_1,...,\alpha_G$ are independent and identically distributed (e.g., the groups are selected by simple random sampling). Second, the group baseline effects α_g are independent of both x_{gi} and v_{gi}.

In general, the true exposures x_{gi} will be unknown, and will be estimated by measured values, as discussed in the following section. For the remainder of this section, we will ignore the effect of measurement error, effectively assuming the true exposures to be known. We will also assume that the true values of the confounder v_{gi} are known, although measurement error in v_{gi} can bias the estimator of β (*3*).

Equation 1 can be used to estimate β, and is appropriate when the α_g's are considered fixed. When the α_g's are independent random variables with $E(\alpha_g = \alpha)$, $\mathrm{var}(\alpha_g) = \tau^2$, an estimator with smaller variance is obtained using the equation

$$y_{gi} = \alpha + \beta x_{gi} + \gamma v_{gi} + \eta_{gi} \qquad [2]$$

The error η_{gi} is equal to $\alpha_g - \alpha + \varepsilon_{gi}$. The covariance matrix of η can be described as follows: Let $\rho = \tau^2/(\sigma^2 + \tau^2)$. Define $\Sigma = (1-\rho)\ \mathbf{I} + \rho \mathbf{1}\mathbf{1}^T$, where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is an I-dimensional column of 1s. Define Σ_{BIG} to be the $GI \times GI$ block diagonal matrix, consisting of G identical blocks of the matrix Σ. Then the covariance matrix of η is equal to $(\sigma^2 + \tau^2)\Sigma_{BIG}$. The matrix Σ^{-1} is equal to $a\mathbf{I} + b\mathbf{1}\mathbf{1}^T$, where $a = 1/(1-\rho)$, and $b = -\rho/\{[(I-1)\rho + 1](1-\rho)\}$. Thus Σ_{BIG}^{-1} is a block diagonal matrix with each block equal to Σ^{-1}.

If ρ is known, the parameters α, β, and γ can be estimated by weighted least squares. If σ^2 and τ^2 are unknown, the parameters can be estimated by a two-stage procedure. In the first stage, only within-groups differences are used. This is accomplished by using Equation 1 to estimate the parameters $\alpha_1,...,\alpha_G$, β, γ by ordinary least squares. Denote by $\hat{\beta}_1$ the estimate of β obtained from this first stage regression, and by $\hat{\sigma}^2$ the usual mean square residual estimate of error variance. The second stage regression involves only the between-groups differences. The regression equation is obtained from Equation 2 by averaging over i: .

$$\bar{y}_{g\cdot} = \alpha + \beta \bar{x}_{g\cdot} + \gamma \bar{v}_{g\cdot} + \bar{\eta}_{g\cdot} \qquad [3]$$

The variables η_g are independent with mean 0 and variance $\tau^2 + \sigma^2 / I$. Denote by $\hat{\beta}_2$ the ordinary least squares estimate of β from Equation 3. The mean square residual is an estimate of $\tau^2 + \sigma^2 / I$, which can be combined with $\hat{\sigma}^2$ to yield an estimate of τ^2. The estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are uncorrelated. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ denote the design matrices from the first and second stages, respectively. Then Var $(\hat{\beta}_1)$ and Var $(\hat{\beta}_2)$ are estimated with appropriate elements from the diagonals of the matrices $(\mathbf{X}_1^T \mathbf{X}_1)^{-1}\hat{\sigma}^2$ and $(\mathbf{X}_2^T \mathbf{X}_2)^{-1}(\hat{\sigma}^2/I + \hat{\tau}^2)$. The two-stage procedure is completed by computing the variance weighted average of $\hat{\beta}_1$ and $\hat{\beta}_2$ to obtain the estimator

$$\hat{\beta}_{pooled} = \frac{\mathrm{Var}\left(\hat{\beta}_2\right)\hat{\beta}_1 + \mathrm{Var}\left(\hat{\beta}_1\right)\hat{\beta}_2}{\mathrm{Var}\left(\hat{\beta}_1\right) + \mathrm{Var}\left(\hat{\beta}_2\right)} \qquad [4]$$

The relationship between weighted least squares and the two-stage procedure is given by the following:

Theorem: Let $\hat{\sigma}^2$, $\hat{\tau}^2$ be the estimators of σ^2, τ^2 from the two-stage procedure. Then $\hat{\beta}_{pooled}$ is the weighted least squares estimate of β when $\rho = \hat{\tau}^2/\hat{\sigma}^2 + \hat{\tau}^2)$.

Corollary: If the errors η_{gi} are normally distributed, then the MLE of β satisfies Equation 4, with $\hat{\beta}_{MLE}$ substituted for $\hat{\beta}_{pooled}$ and Var $(\hat{\beta}_1)$ and Var $(\hat{\beta}_2)$ evaluated at the MLEs of σ^2 and τ^2.

Proofs of these claims are provided in the Appendix. The corollary suggests that Equation 4, if iterated, will converge to the MLE.

Allowance for Exposure Measurement Error

In many circumstances, it may not be feasible to obtain complete and error-free data on all individuals, and hence some variables will only be available for some (randomly selected) subset of individuals. For example, in a dietary study, one might wish to validate the use of a food frequency questionnaire in the entire group by repeated 7-day records. In an air pollution study, it might be feasible to obtain personal monitoring or microenvironment sampling data on only a sample, but questionnaire data on individual modifying factors might be available on the entire group. Optimization of the design typically would entail trade-offs between the number of groups and the number of individuals in the main study and in validation substudies, and the extent of the measurement protocols, subject to constraints on the total costs. These design issues will be discussed further below. In this section, we will focus on the effect of exposure measurement error. To simplify matters, we will ignore confounding.

We make a distinction in our analysis between two types of measurement error. The first type, known as the "Berkson" error model (*4*), applies when individuals are assigned their group average exposures. The second type, known as the "classical" error model, applies when the assigned exposure is a random variable whose expected value is the true exposure.

Let x_{gi} denote the unobservable true exposure for individual i in group g and let z_{gi} indicate the measured value (e.g., from personal monitoring). The classical error model assumes that the measured values are randomly distributed around the true value with the property that $E(z_{gi} | x_{gi}) = x_{gi}$. As is well known [reviewed recently by Thomas et al. (*5*)], the classical error model produces a bias towards the null, essentially because the measured exposures are overdispersed ($\mathrm{Var}(z_{gi}) = \mathrm{Var}(x_{gi}) + \mathrm{Var}(z_{gi} | x_{gi}) > \mathrm{Var}(x_{gi})$). Thus if $\mathrm{Var}(x_{gi}) = \phi_g^2$ and $\mathrm{Var}(z_{gi} | x_{gi}) = \phi^2$, the regression on z_{gi} produces a slope estimate $\hat{\beta}$ that has expectation $c_g = \phi_g^2/(\phi_g^2 + \omega^2)$ times the expectation of the slope of the regression on the x_{gi}. This suggests a simple correction for measurement error if these variances are equal and known. First fit the naive regression on z_{gi} and then correct the estimated slope coefficient by dividing it by c (*6*). For more complex situations, for example if the variances differ between groups, a useful strategy is to replace the z_{gi}'s by $\hat{x}_{gi} = \mathrm{E}(x_{gi} | z_{gi}) = c_g z_{gi} + (1-c_g)\mathrm{E}(x_{gi})$ and then use these $\hat{x}_{gi}$'s as if they were the true exposures in the regression.

The Berkson error model assumes instead that the true exposures x_{gi} of individuals are distributed around their group estimates X_g with the property that $E(x_{gi} | X_g) = X_g$. Thus, in an air pollution study with no personal monitoring, we might assume that individuals' exposures are randomly distributed around the ambient levels for their communities. A consequence of this assumption is that, at least for linear dose-response models, the regression on the measured values provides unbiased estimates of the true slope. If $y_{gi} = \alpha_g + \beta x_{gi} + \varepsilon$, then

$$E(y_{gi} | X_g) = \alpha_g + \beta \mathrm{E}(x_{gi} | X_g) + E(\varepsilon | X_g)$$
$$= \alpha_g + \beta X_g.$$

Thus, Berkson error produces no bias towards the null for linear models.

Typically, it would not be feasible to obtain true exposure data on any individuals. Rather, a surrogate variable w would be obtained on everybody and higher quality measurements z only on a sample. The measurements are assumed to be unbiased in the classical error sense and might be replicated T times. In this case, it will not be possible to use the z's directly in modeling y because they are available on too few subjects; but they could be used to build a model for the relationship between z and w, which could be then be used for imputing $\hat{x}$ values in the first stage regression. The surrogate variable w might be a simpler measure of x (such as a food frequency questionnaire) or it might be a personal modifier of a group exposure characteristic X (for example, percent time spend outdoors in an air pollution study could modify the ambient pollution level).

To give a concrete example of this imputation procedure, assume that at times $t = 1,2,\ldots,T$ we have measurements of a group exposure characteristic X_{gt} for each group, and for a subset of individuals we have an exposure modifying variable u_{git} and an exposure measurement z_{git}. We assume that X and w are assessed without error, and z has a classical error structure in relation to true exposure x. We assume the following relationships:

$$x_{git} \sim N(X_{gt} + \delta_0 + \delta_1 w_{git}, \phi^2) \quad [5]$$

$$z_{git} \sim N(x_{git}, \omega^2) \quad [6]$$

We assume that ω^2 is known from other studies or from another set of replicate measurements, but that δ_0, δ_1, and ϕ^2 are unknown. Combining Equations 5 and 6 yields

$$z_{git} \sim N(X_{gt} + \delta_0 + \delta_1 w_{git}, \phi^2 + \omega^2) \quad [7]$$

from which we can obtain unbiased estimates of δ_0, δ_1, and ϕ^2 (since ω^2 is known). We then estimate x_{gi} as $\hat{x}_{gi} = X_g + \hat{\delta}_0 + \hat{\delta}_1 w_{gi}$, which is an unbiased estimator of $E(x_{gi} | X_g, w_{gi})$ since $\hat{\delta}_0$ and $\hat{\delta}_1$ are unbiased.

Allowing for measurement error complicates the two-stage procedure for estimating the parameter β as follows. We assume

$$\alpha_g \sim N(\alpha, \tau^2) \quad [8]$$

$$y_{gi} \sim N(\alpha_g + \beta x_{gi}, \sigma^2) \quad [9]$$

where α, τ^2, and σ^2 are unknown. Since the x_{gi} are also unknown, we replace them with their estimates $\hat{x}_{gi}$ when fitting the model. The first stage model is thus:

$$y_{gi} = \alpha_g + \beta \hat{x}_{gi} + \varepsilon_{gi}, \quad [10]$$

where the ε_{gi} are independent normal random variables with $\mathrm{E}(\varepsilon_{gi}) = 0$ and $\mathrm{Var}(\varepsilon_{gi}) = \sigma^2 + \beta^2 \mathrm{E}[(x_{gi} - \hat{x}_{gi})^2 | X_g, w_{gi}]$.

Since $\mathrm{Var}(\varepsilon_{gi})$ depends on g, i, and the unknown parameter β, the model, Equation 10, may be fit by iteratively reweighted least squares (IRLS). To use this procedure, we must first express the weights $\mathrm{Var}(\varepsilon_{gi})$ in usable form. Let $\hat{V}(\hat{\delta}_0)$, $\hat{V}(\hat{\delta}_1)$, and $\hat{C}(\hat{\delta}_0, \hat{\delta}_1)$, be the estimates of $\mathrm{Var}(\hat{\delta}_0 | X_g, \{w_{gi}\})$, $\mathrm{Var}(\hat{\delta}_1 | X_g, \{w_{gi}\})$, and $\mathrm{Cov}(\hat{\delta}_0, \hat{\delta}_1 | X_g, \{w_{gi}\})$, respectively, calculated from the regression model, Equation 7. Since $\hat{x}_{gi} \approx \mathrm{E}(x_{gi} | X_g, u_{gi})$, $\mathrm{Var}(\varepsilon_{gi})$ can be approximated by

$$V^* = \sigma^2 + \beta^2 \mathrm{Var}(x_{gi} | X_g, w_{gi}) \approx \sigma^2 + \beta^2 W_{gi}$$

where

$$W_{gi} = \hat{V}(\hat{\delta}_0) + 2 w_{gi} \hat{C}(\hat{\delta}_0, \hat{\delta}_1) + w_{gi}^2 \hat{V}(\hat{\delta}_1) + \hat{\phi}^2.$$

The IRLS procedure is then conducted as follows. Set $\hat{\sigma}^2$ and $\hat{\beta}$ to arbitrary initial values, then fit model Equation 10 by weighted least squares regression using weights V^{*-1}. This produces an updated estimate $\hat{\beta}^{(1)}$ of β, and fitted values $\hat{y}_{gi}^{(1)}$. To obtain an updated estimate of σ^2, take the average of the values $(y_{gi} - \hat{y}_{gi}^{(1)})^2 - (\hat{\beta}^{(1)})^2 W_{gi}$ and then repeat the entire process.

An alternative to IRLS is maximum likelihood. The MLEs of α_g, β, and σ^2 can be obtained by minimizing

$$\sum_{g=1}^{G}\sum_{i=1}^{I}\sum_{t=1}^{T}\left[\log(\phi^2+\omega^2) + \frac{(z_{git}-X_{gt}-\delta_0-\delta_1 w_{git})^2}{2(\phi^2+\omega^2)}\right] + \sum_{g=1}^{G}\sum_{i=1}^{I}\left[\log(\sigma^2+\beta^2\phi^2) + \frac{(y_{gi}-\alpha_g-\beta\delta_0-\beta\delta_1 w_{gi})^2}{2(\sigma^2+\beta^2\phi^2)}\right] \quad [11]$$

over values of α_g, $g=1,\ldots,G$, β, δ_0, δ_1, ϕ, and σ^2.

Let $\hat{\alpha}_g$, $\hat{\beta}_1$, and σ^2 be the estimates of α_g, β, and σ^2 obtained from the first-stage model Equation 10. Let $\mathbf{W}$ be the diagonal matrix whose diagonal elements are $\hat{\sigma}^2+\hat{\beta}_1 W_{gi}$. Let $\mathbf{X}$ be the design matrix corresponding to the first-stage model Equation 10. The usual estimate of the covariance matrix of $\hat{\alpha}_g$, $\hat{\beta}_1$, is $(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}$, which does not seem to be accurate when the exposure is poorly estimated (see simulation results below). An alternative which may prove more satisfactory is the "sandwich estimator" (*7*), which may be less sensitive to misspecification or variation in the weight matrix $\mathbf{W}$. Other alternatives include computing the MLE of $\hat{\beta}_1$ using Equation 11 and using the estimated Fisher information, or estimating the variance with a bootstrap procedure. All the estimators just described provide estimates of the conditional variance of $\hat{\beta}$ given X_g and w_{gi}. They will serve as estimates of the unconditional variance with no additional bias if the conditional expectation of $\hat{\beta}$ given X_g and w_{gi} is constant across values of X_g and w_{gi}.

The second-stage model is obtained from Equation 10 by averaging over i and by replacing α_g with its mean α to obtain

$$\bar{y}_{g\cdot} = \alpha + \beta\hat{x}_{g\cdot} + \eta_g \quad [12]$$

where $\hat{x}_{g\cdot} = I^{-1}\,\Sigma_{i=1}^{I}\,\hat{x}_{gi}$, and η_g are independent random variables with $\mathrm{E}(\eta_g)=0$ and

$$\mathrm{Var}(\eta_g)=\tau^2+\sigma^2/I+\beta^2\mathrm{E}[(\overline{x}_{\cdot}-\hat{x}_{g\cdot})^2 | X_g,\{w_{gi}\}]. \quad [13]$$

The expectation on the right hand side of Equation 13 may be estimated with $\overline{W_{g\cdot}}$, the average of the W_{gi} described above.

The second-stage model may also be fitted by IRLS, updating β as in the first-stage regression, and updating the sum $\tau^2+\sigma^2/I$ as a single quantity. After convergence, a separate estimate of τ^2 can be obtained by combining the estimate of $\tau^2+\sigma^2/I$ with the estimate of σ^2 from the first-stage regression. Let $\hat{\beta}_2$ be the estimate of β from the second-stage regression. The covariance matrix of $\hat{\alpha}$ and $\hat{\beta}_2$ can be obtained in a manner analogous to one of the methods described above for the first-stage model.

If the true weights were known and did not depend on β, the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ would be uncorrelated. When variables are normally distributed, they are nearly uncorrelated when the weights are estimated as well (simulation results, not reported). Preliminary calculations seem to indicate that in some cases measurement error may introduce a substantial correlation. Unless this correlation can be estimated, it is not clear which linear combination of the two estimates is optimal. In particular, it is difficult to judge the performance of the variance-weighted estimator $\hat{\beta}_{\text{pooled}}$ given by Equation 4.

For many applications in environmental epidemiology, it is more appropriate to assume that true exposures are nonnegative and lognormally distributed and that measurement errors are lognormally distributed and multiplicative. Furthermore, the individual exposure modifiers w_{gi} might also be assumed to act multiplicatively on the group means X_g. All this can be accomplished without any new theory simply by redefining x, X, and z to be the logarithms of their respective quantities. For chronic exposures, however, it may be more appropriate to relate the outcome y to the time-weighted-average (arithmetic mean) or cumulative exposure than to the geometric mean or integral of the log exposures. This leads to additional complexities involving means and variances of lognormal distributions.

Compatibility of First- and Second-stage Models

If the assumption that group baseline effects are independent of group mean exposure levels (assumption 2) is violated, the second-stage model Equation 12 may produce a biased estimate of β. This is the case because the error term η_g in Equation 12 is correlated with the group baseline effect α_g. This result is precisely the ecologic fallacy, wherein effects due to variation in α_g appear to be explained by variation in $\bar{x}_g$. It is therefore wise to test for bias in $\hat{\beta}_2$ before pooling it with $\hat{\beta}_1$. One way to do this is to test whether the difference $\hat{\beta}_1-\hat{\beta}_2$ is significantly different from 0. In the absence of measurement error, the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ will be nearly uncorrelated even if α_g and $\bar{x}_g$ are dependent, so the variance of the difference can be estimated by adding the variance estimates from the first and second stage regressions. In the presence of measurement error, it is not yet clear how to estimate this variance accurately.

Design Optimization

At the design stage, the epidemiologist needs to consider the trade-off between the number of groups and the number of subjects per group, the selection of the specific groups to be included, the number of subjects in the main study versus the number in the validation sample, and the number and complexity of measurements to be made on each sample. These are important issues that have been given only limited attention in the context of analytic studies and none in the context of ecologic or hybrid designs. For analytic studies, Greenland (*8*) and Spiegelman and Gray (*9*) have considered the trade-offs between numbers of subjects in the main and validation studies and provided explicit formulas for determining the optimal design where it is planned to use measurement error adjustment methods in the analysis like those described above. Rosner and Willett (*10*) considered the trade-off between numbers of subjects and numbers of replicate measurements in a validation study.

For linear models with a continuous normally distributed outcome, ignoring confounding at the individual and group levels, measurement error, and assuming exposure is assessed only at the group level, the power of the study can be computed as a function of four quantities: the number of groups G, the number of subjects I sampled in each group, the true R^2 between group mean exposure and group mean outcome, and the ratio $VR=V_W/V_B$, where V_W and V_B are outcome variances within and between groups respectively. Given these quantities, we compute $R_*^2=IV_BR^2/(V_W+IV_B)$, the squared correlation between group mean exposure and the average outcome among the individuals sampled from the group. The quantity R_*^2 is less than R^2, because the sample mean outcome rather than the true group mean outcome is used. The power to detect a nonzero R^2 is calculated by using Fisher's transformation of R_*^2.

Table 1 illustrates the results for a variety of choices of the model and design parameters. It is clear that the power is much more strongly influenced by the number of groups than by the number of subjects. For a logistic model for binary

Table 1. Statistical power of between-groups comparisons.

		R^2 for between-groups regression					
		R^2= 0.1		R^2= 0.3		R^2= 0.5	
G	$G \times I$	VR=10	VR=100	VR=10	VR=100	VR=10	VR=100
5	1000	0.12	0.10	0.21	0.18	0.33	0.28
	4000	0.12	0.11	0.22	0.20	0.34	0.32
10	1000	0.21	0.15	0.46	0.32	0.72	0.54
	4000	0.21	0.19	0.48	0.43	0.75	0.68
20	1000	0.34	0.20	0.75	0.45	0.96	0.73
	4000	0.37	0.29	0.80	0.68	0.97	0.92
40	1000	0.52	0.23	0.94	0.54	1.00	0.84
	4000	0.60	0.41	0.97	0.86	1.00	0.99

G, number of groups; I, number of individuals per group; $VR = V_W/V_B$, where V_W is the within groups outcome variance and V_B is the between groups outcome variance.

Table 2. Statistical power of between-individuals (within groups) comparisons.

$G \times I$	R^2= 0.001	R^2= 0.005	R^2= 0.01	R^2= 0.05
1000	0.26	0.72	0.94	1.00
2000	0.41	0.94	1.00	1.00
4000	0.64	0.99	1.00	1.00

outcomes, the power also depends on the overall disease frequency, but the same basic result emerges—the power of the aggregate analysis depends much more strongly on the number of groups than on the number of subjects per group.

Table 2 provides similar power calculations for testing a partial R^2 for the individual regression after removing group effects, again using Fisher's transformation. The power of these analyses depends only on the total number of individuals and it is clear that with sample sizes in the thousands, there is adequate power for detecting very small correlations. However, it is important to note that these are the correlations with the measured exposures, which could be severely attenuated by measurement error.

To provide further guidance for the design of the USC air pollution study, we undertook a limited simulation study. For this purpose, we varied the number of groups G, the number of subjects per group in the main study I, and the number of subjects per group in the exposure substudy S. Relationships among the variables were as given in Equations 5 to 9, with the ambient levels X_g, X_{gt} and individual modifiers w_{gi}, w_{git} being normally distributed. For each choice of design parameters, 1000 replicate data sets were simulated and analyzed using the methods described above. We tabulated the bias and variance of the parameter estimates from the individual level regressions (with and without adjustment for measurement error), the ecologic regression, and the proposed pooled combination of the two regressions.

The design parameters were chosen to approximate those being considered for the USC air pollution study, and the model parameters were then adjusted to illustrate a hypothetical situation in which the two approaches to estimation would be roughly equally informative. Table 3 illustrates the effect of modifying the design parameters under the constraint that the total number of measurements $G(I+SM)$ be fixed at 3000. (A more realistic simulation would allow for differences in costs between the different types of measurements.) Under the assumptions of the simulation, measurement error is minimized when one measurement is taken per individual in the substudy. Therefore we set $T = 1$. All parameter estimates appear to be nearly unbiased. The columns labeled "sample standard error (SE)" give the sample standard deviation of the 1000 parameter estimates. The columns labeled "nominal SE" give the square root of the average of the 1000 conditional variance estimates obtained from the covariance matrix estimator $(X^TW^{-1}X)^{-1}$. Each block of the table shows the effect of varying the number of groups and the number of subjects per group. In agreement with Tables 1 and 2, the efficiency of $\hat{\beta}_2$ improves rapidly as the number of groups increases, whereas $\hat{\beta}_1$ depends only on the total number of subjects. Comparing the two blocks illustrates the trade off between the number of subjects in the main study and validation substudy. The second-stage estimator is relatively insensitive to this parameter, while the first-stage estimator is improved by having a larger proportion in the validation study, although we do not have enough information to determine the optimal allocation. Perhaps more important, when too few subjects are assigned to the substudy, the nominal SE of $\hat{\beta}_1$ is far too optimistic, since it fails to take into account the error in misspecifying the weights. Since $\hat{\beta}_{\text{pooled}}$ is based on the nominal SEs, it is no longer the optimal linear combination of $\hat{\beta}_1$ and $\hat{\beta}_2$, and in some cases is less efficient than $\hat{\beta}_2$.

Table 3. Standard errors of parameter estimates in the presence of measurement error.

			SE ($\hat{\beta}_1$)		SE ($\hat{\beta}_2$)		SE ($\hat{\beta}_{pooled}$)	
G	I	S	Sample SE	Nominal SE	Sample SE	Nominal SE	Sample SE	Nominal SE
6	200	300	0.19	0.15	0.32	0.30	0.15	0.13
12	100	150	0.19	0.15	0.19	0.19	0.13	0.11
24	50	75	0.19	0.15	0.14	0.14	0.11	0.10
48	25	38	0.18	0.15	0.11	0.11	0.092	0.086
6	400	100	0.24	0.11	0.30	0.30	0.19	0.10
12	200	50	0.25	0.11	0.18	0.18	0.16	0.090
24	100	25	0.24	0.11	0.12	0.12	0.12	0.080
48	50	13	0.23	0.11	0.092	0.092	0.10	0.069

G, number of groups; I, number of individuals per group in the main study; S, number of individuals per group in the exposure validation substudy. 1000 data sets were generated for each value of G, I, and S. Exposures measured subject to error; measurement error variance estimated in substudy. True parameter values are $\phi^2 = 1.0$, $\omega^2 = 25.0$, $\tau^2 = 1.0$, $\sigma^2 = 25.0$. In addition, $\text{Var}(X_g) = 4.0$, $\text{Var}(X_{gt} | X_g) = 0.25$, $\text{Var}(w_{gi}) = 1.0$, and $\text{Var}(w_{git} | w_{gi}) = 0.25$.

Example: The USC Air Pollution Study

In January 1992, the California Air Resources Board (ARB) awarded a contract to the University of Southern California to initiate a 10-year cohort study of the health effects of air pollution in southern California. The study will enroll a cohort of about 3500 school children from 12 communities selected to represent a variety of types and levels of air pollution that are represented in the basin. The primary focus of the study is on the effects of chronic exposure to 1-hr peak ozone (O_3), but particulates (PM_{10}), nitrogen dioxide (NO_2), acids (H^+), and other pollutants are also being measured. Health outcomes to be measured annually will include various lung function tests, symptoms reported by questionnaire, and absences abstracted from school records.

Table 4. Characteristics of the communities selected for the southern California air pollution study.

		Annual mean level[a]				Demographic characteristics		
Community	Profile[b]	O_3	PM_{10}	NO_2	H^+	% White	% Age 5–18	People/room
Glendora	++++	109.2	67.0	39.1	2.93	89	20	0.50
Upland	++++	92.0	75.6	44.6	3.09	75	22	0.58
Rubidoux	+++–	95.1	84.9	32.8	1.05	68	24	0.62
Riverside	+++–	95.1	84.9	32.8	1.05	77	22	0.51
Perris	++–+	81.3	60.1	15.4	1.99	73	23	0.62
Lancaster	++–+	70.8	47.0	13.2	3.16	81	21	0.54
Lake Gregory	+–++	98.8	38.3	23.6	2.26	95	22	0.51
Alpine	+–––	80.5	37.4	16.7	1.18	93	19	0.48
North Long Beach	–+++	45.2	49.5	44.8	2.43	58	18	0.59
Santa Maria	––––	30.2	28.0	7.7	0.91	66	20	0.61
Santa Barbara	––––	30.4	31.0	10.4	0.91	84	19	0.50
Lompoc	––––	34.8	30.0	1.6	0.91	72	21	0.58

[a] O_3 and NO_2 are measured in parts per billion on a mass basis. PM_{10} is measured in micrograms per cubic meter. H^+ is measured in parts per billion on a mole basis. [b] + signifies that the pollution level is above the mean level of the 86 communities considered, – signifies that the pollution is below that level.

Table 5. Comparison of power to detect effects of four priority pollutants from alternative choices of sites.

									Power			
Community selection based on:				Pollutants included in model				G^a	O_3	PM_{10}	NO_2	H^+
O_3	PM_{10}		H^+	O_3	PM_{10}	NO_2	H^+	12	0.88	0.66	0.23	0.28
O_3	PM_{10}	NO_2		O_3	PM_{10}	NO_2	H^+	12	0.89	0.89	0.40	0.30
O_3	PM_{10}	NO_2	H^+	O_3	PM_{10}	NO_2	H^+	12	0.78	0.87	0.58	0.77
Random				O_3	PM_{10}	NO_2	H^+	12	0.34	0.61	0.34	0.37
O_3	PM_{10}			O_3	PM_{10}	NO_2	H^+	8	0.48	0.33	0.14	0.14
O_3	PM_{10}	NO_2		O_3	PM_{10}	NO_2	H^+	6	0.07	0.08	0.06	0.06
O_3	PM_{10}		H^+	O_3	PM_{10}	NO_2	H^+	6	0.13	0.10	0.07	0.07
Original				O_3	PM_{10}			5	0.22	0.23		

[a] Number of groups in the study.

Community Selection

Some preliminary power calculations based on assumed values for true effects indicated that for studying a single pollutant, it would be necessary to have at least ten groups for power to be adequate. We carried out further calculations along similar lines to assess the prospects for doing multivariate analyses of two or more pollutants and concluded that it would be possible, provided groups could be selected in such a way that the correlations in pollutant levels across groups were not too large. Thus, the optimal choice would have to take account of the actual levels of exposure to each of the pollutants we wished to assess.

Fortunately, extensive data were available on the four highest priority pollutants from the ARB's monitoring program. Year-round average levels for the period 1986 to 1990 were obtained from 86 monitoring stations scattered across southern California. (For some pollutants, notably acids, the values had to be interpolated from other stations on an inverse-distance weighted basis). Our initial selection of sites was based on the intuitive notions that we wished to maximize the dispersion of each of the pollutants, and we wished to represent as many combinations of high and low levels of each pollutant as possible. These notions are appropriate when the response surface is linear.

For each pollutant, we calculated the mean level over the 86 communities, then for each community, we converted the pollution levels to standard units. Each community was assigned a "profile" by recording it as either above (+) or below (–) the mean level for each pollutant. For a design based on all four pollutants, there were thus $2^4 = 16$ possible profiles, of which demographically suitable examples could be found for seven of them. Within each profile, we then selected from one to three communities whose sum of squared standardized pollution levels were large. Table 4 describes the characteristics of the communities that we judged to be the most suitable on this basis, under the constraint that we could afford to study no more than 12. This selection process differs from the one described above in that the groups were not randomly chosen. Thus the group effects must be considered fixed rather than random.

To compare alternative designs based on different selections of priority pollutants, we then carried out a further simulation study, based purely on the second stage ecologic regression but allowing the actual pollutant levels to differ from the measured values subject to a covariance structure estimated from the observed data [detailed in Peters (*11*)]. Table 5 summarizes the results of this simulation, which led us to the conclusion that, if all four pollutants had health effects, then the optimal design would need to be based on all four. This design appears to have adequate power for detecting differences in mean forced expiratory volume in one second (FEV) of about 3 to 5% between the high and low communities for each of the four pollutants in multivariate analysis, assuming that one-third of the variance in FEV_1 is explained by variation in the pollutants, and that O_3 and PM_{10} each contribute twice as much to the health effect as do NO_2 and H^+. Alternative designs that ignore one or more of these pollutants (with the same total number of communities) may slightly increase the variability of the pollutants of primary interest, which normally would be expected to yield an improvement in power. However, they also substantially weaken the power for controlling the confounding effect of the omitted pollutants and therefore in most instances reduce the power for the effects of interest in a multivariate analysis.

To determine whether we could significantly improve our selection of communities under the four-pollutant design, we conducted a final simulation along similar lines, starting with the choice given in Table 4 and in a stepwise fashion considered replacing each of the 12 communities by each of the remaining candidates. This led to the conclusion that, under an optimality criterion that maximized the sum of the powers for the four pollutants, it was theoretically possible to improve the design further by changing 5 of the 12 sites. This alternative choice attained better overall power by substantially reducing the correlations among the exposure variables. However, it did so at the expense of substantially reducing the variance of each exposure. Since we were unsure of the validity of the correlation estimates because many of the entries were based on interpolation, and since the overall improvement in power was modest, we decided to retain our original selection. Essentially, we judged that the primary objective of the study was to maximize the overall power to detect any air pollution effect and that the separation of the effects of particular pollutants was of only secondary importance, after having demonstrated an overall effect. We therefore felt that it was more important to maximize the variance in exposures than to minimize their covariances.

Exposure Modeling

The measurement protocol entails a combination of ambient monitoring, personal monitoring, microenvironment sampling, and questionnaire assessment of personal modifying factors. Ambient data are routinely collected by the ARB for each of the communities, and will provide long-term average levels throughout the study as well as historically. The questionnaire will be administered to all subjects and will include items on residence history, usual indoor and outdoor times and activities, and household characteristics (smoking by family members, air conditioning and heating, air exchange, sources of indoor pollution, etc.). Personal monitoring will be possible only for ozone and only on a sample of subjects. These subjects will also maintain a daily diary of their activities during the times when the monitoring badge is worn. Microenvironment sampling will be done on all pollutants at a variety of indoor and outdoor locations in each community.

The goal of the analysis will be to combine these various data sources in such a way as to provide estimates of individual and group mean exposures for the first- and second-stage regressions described above, including estimates of measurement error distributions for adjustment purposes. The actual form of the models to be used is still under development, and will incorporate the extensive body of literature on the determinants of personal exposure. To illustrate the general approach, we make some simplifying assumptions that will be remedied in our final analyses.

First, we assume that the relevant exposure variable is the long-term arithmetic mean (i.e, the "time-weighted average," TWA). We also assume that ambient levels, true personal exposures, and measurement errors are lognormally distributed. Finally, we assume that the ratio of personal exposures to ambient levels is described by a multiplicative factor that depends loglinearly on the personal modifying factors. The basic relationships are thus as described in Equations 5 to 10, except for the additional complexities introduced by the lognormal assumptions. Using the estimates from this model, we can compute for each subject in the main study the TWA, $E(e^{x_{gi}}|X_g,w_{gi})$, for use in the first-stage regression, together with the average over all subjects of these TWAs for use in the second-stage regression. Whether it will be possible to assess exposure effects at an individual level will depend primarily on the variability between individuals in their modifying factors and the ability of the exposure model to accurately predict personal exposures. Even if it is not possible to assess dose-response relations at an individual level, however, the use of average TWAs rather than X_g in the second stage should lead to more reliable estimates, because communities with different exposure patterns are likely to differ substantially in modifying factors such as use of air conditioning and proportion of time spent outdoors, because of major differences in climate across southern California.

Conclusions

It is reasonable to expect that the proposed two-stage analysis of the multilevel analytic design will provide unbiased and efficient estimation of effects in a complex model involving unmeasured between-group differences, measurement error, and a complex measurement model combining individual and aggregate exposure data. In particular, in cases where the within-groups exposure variance is less than the between-groups variance, estimates obtained through pooling should be more efficient than estimates based on either individual level or aggregate level analyses alone. Simulation techniques can be used to optimize the various trade-offs between the design parameters if reasonable estimates of the model parameters are available. We believe this design and its associated analysis offer considerable promise for resolving some of the difficulties of between-group confounding, measurement error, and restricted variability that have historically plagued environmental epidemiology.

APPENDIX

Proof of Theorem: We prove the theorem for a more general case with an arbitrary number of confounders. The model is

$$y_{gi}=\alpha+\beta x_{gi}+\mathbf{V}\gamma+\eta_{gi}$$

where $\mathbf{V}$ is a matrix of confounder variables. Let $\hat{\sigma}^2$ and $\hat{\tau}^2$ be estimates of σ^2 and τ^2, and let $\hat{\Sigma}_{BIG}$ be the corresponding estimate of Σ_{BIG}.

Assume without loss of generality that $\hat{\Sigma}^{-1/2}_{BIG}\mathbf{x}$ is orthogonal to $\hat{\Sigma}^{1/2}_{BIG}\mathbf{V}$. Otherwise, replace $\mathbf{x}$ with $\mathbf{x}-\mathbf{V}(\mathbf{V}^T\hat{\Sigma}^{-1}_{BIG}\mathbf{V})^{-1}\mathbf{V}^T\hat{\Sigma}^{-1}_{BIG}\mathbf{x}$ and reparameterize the confounders γ. The weighted least squares estimates of α and β are the values minimizing

$$\mathbf{y}-\alpha-\beta\mathbf{x})^T\hat{\Sigma}^{-1}_{BIG}(\mathbf{y}-\alpha-\beta\mathbf{x}) \qquad [14]$$

which is equal to

$$\sum_{g=1}^{G}(\mathbf{y}_g-\alpha-\beta\mathbf{x}_g)^T\hat{\Sigma}^{-1}(\mathbf{y}_g-\alpha-\beta\mathbf{x}_g) \qquad [15]$$

where $\mathbf{y}_g=(\mathbf{y}_{g1},\ldots,\mathbf{y}_{gI})^T$, and $\mathbf{x}_g$ is defined similarly.

Let $\rho=\hat{\tau}^2/(\sigma+\tau)$, $a=1/(1-\rho)$, and $b=-\rho/(1-\rho)(1+(I-1)\rho)$. Substitute $(a\mathbf{I}+b\mathbf{1}\mathbf{1}^T)/(\hat{\sigma}^2+\hat{\tau}^2)$ for $\hat{\Sigma}^{-1}$ in Equation 15 to obtain

$$\frac{1}{\hat{\sigma}^2+\hat{\tau}^2}\sum_{g=1}^{G}\begin{bmatrix} a\sum_{i=1}^{I}(y_{gi}-\alpha-\beta x_{gi})^2+ \\ bI^2(\bar{y}_{g\cdot}-\alpha-\beta\bar{x}_{g\cdot})^2 \end{bmatrix} \qquad [16]$$

Algebraic manipulation yields

$$\frac{a}{\hat{\sigma}^2+\hat{\tau}^2}\sum_{g=1}^{G}\sum_{i=1}^{I}\left[y_{gi}-\bar{y}_{g\cdot}-\beta(x_{gi}-\bar{x}_{g\cdot})\right]^2+\frac{Ia+I^2b}{\hat{\sigma}^2+\hat{\tau}^2}\sum_{g=1}^{G}\left[\bar{y}_{g\cdot}-\alpha-\beta\bar{x}_{g\cdot}\right]^2. \qquad [17]$$

Substitute appropriate expressions for a and b in terms of $\hat{\sigma}^2$ and $\hat{\tau}^2$ to obtain

$$\frac{1}{\hat{\sigma}^2}\sum_{g=1}^{G}\sum_{i=1}^{I}\left[y_{gi}-\bar{y}_{g\cdot}-\beta(x_{gi}-\bar{x}_{g\cdot})\right]^2+\frac{1}{\hat{\sigma}^2/I+\hat{\tau}^2}\sum_{g=1}^{G}\left[\bar{y}_{g\cdot}-\alpha-\beta\bar{x}_{g\cdot}\right]^2. \qquad [18]$$

Differentiating with respect to α and β and setting partial derivatives equal to 0 shows that the value of β minimizing Equation 18 is

$$\beta_{WLS}= \frac{\Sigma_g\Sigma_i(x_{gi}-\bar{x}_{g\cdot})(y_{gi}-\bar{y}_{g\cdot})/\hat{\sigma}^2 +}{\Sigma_g\Sigma_i(x_{gi}-\bar{x}_{g\cdot})^2/\hat{\sigma}^2 +}$$

$$\frac{\Sigma_g(\bar{x}_{g\cdot}-\bar{x}_{..})(\bar{y}_{g\cdot}-\bar{y}_{..})/(\hat{\sigma}^2/I+\hat{\tau}^2)}{\Sigma_g(\bar{x}_{g\cdot}-\bar{x}_{..})^2/(\hat{\sigma}^2/I+\hat{\tau}^2)} \qquad [19]$$

Let $\hat{V}\hat{\beta}_1=\hat{\sigma}^2/\Sigma_g\Sigma_i(x_{gi}-\bar{x}_{g\cdot})^2$ be the estimate of $\hat{\beta}_1$ and let $\hat{V}\hat{\beta}_2=(\hat{\sigma}^2/I+\hat{\tau}^2)/\Sigma_g(\bar{x}_{g\cdot}-\bar{x}_{..})^2$ be the estimate of $\hat{\beta}_2$ from the first and second stage models, respectively.

Multiplying numerator and denominator of Equation 9 by $\hat{V}\hat{\beta}_1 \hat{V}\hat{\beta}_2$ shows that

$$\hat{\beta}_{WLS}=\frac{\hat{V}_{\hat{\beta}_2}\hat{\beta}_1+\hat{V}_{\hat{\beta}_1}\hat{\beta}_2}{\hat{V}_{\hat{\beta}_1}+\hat{V}_{\hat{\beta}_2}}=\hat{\beta}_{pooled}$$

Proof of Corollary: If $\eta \sim N(0, \Sigma_{BIG}^{-1})$, then the MLEs of α, β, σ, and τ are the values minimizing

$$L(\alpha, \beta, \sigma, \tau)=\log(\det(\Sigma_{BIG}))+(\mathbf{y}-\alpha-\beta\mathbf{x})^T\Sigma_{BIG}^{-1}(\mathbf{y}-\alpha-\beta\mathbf{x}). \quad [20]$$

Let $\nu^2=\sigma^2/I+\tau^2$. The value of $\det(\Sigma_{BIG})$ is $[(\sigma^2)^{I-1}\nu^2]G$. Now

$$\begin{aligned} L(\alpha,\beta,\sigma,\tau) &= L(\alpha,\beta,\sigma,\nu)\\ &=G(I-1)\log(\sigma^2)+\frac{1}{\sigma^2}\Sigma_{g=1}^G\Sigma_{i=1}^I[y_{gi}-\bar{y}_{g\cdot}-\beta(x_{gi}-\bar{x}_{g\cdot})]^2\\ &+G\log(\sigma^2)+\frac{1}{\upsilon^2}\Sigma_{g=1}^G[y_{g\cdot}-\alpha-\beta\bar{x}_{g\cdot}]^2. \end{aligned} \qquad [21]$$

For any given value of β, the values of α, β, σ, and ν minimizing L are

$$\hat{\alpha}=\bar{y}_{..}-\beta\bar{x}_{..}$$

$$\hat{\sigma}^2=\frac{\Sigma_{g=1}^G\Sigma_{i=1}^I[y_{gi}-\bar{y}_{g\cdot}-\beta(x_{gi}-\bar{x}_{g\cdot})]^2}{G(I-1)}$$

$$\hat{\upsilon}^2=\frac{\Sigma_{g=1}^G[\bar{y}_{g\cdot}-\bar{y}_{..}-\beta(\bar{x}_{g\cdot}-\bar{x}_{..})]^2}{G}$$

For given values of σ^2, ν^2, it follows from the theorem that the value of β minimizing L is

$$\hat{\beta}=\frac{\Sigma_g\Sigma_i(x_{gi}-\bar{x}_{g\cdot})(y_{gi}-\bar{y}_{g\cdot})/\sigma^2+\Sigma_g(\bar{x}_{g\cdot}-\bar{x}_{..})(\bar{y}_{g\cdot}-\bar{y}_{..})/\nu^2}{\Sigma_g\Sigma_i(x_{gi}-\bar{x}_{g\cdot})^2/\sigma^2+\Sigma_g(\bar{x}_{g\cdot}-\bar{x}_{..})^2/\nu^2}$$

which is Equation 4.

REFERENCES

1. Selvin HC. Durkheim's suicide and problems of empirical research. Am J Sociol 63:607–619 (1958).
2. Prentice RL, Sheppard L. Dietary fat and cancer: consistency of the epidemiologic data, and disease prevention that may follow from a practical reduction in fat consumption. Cancer Causes Control 1:81–97 (1990).
3. Greenland S. The effect of misclassification in the presence of covariates. Am J Epidemiol 112:564–569 (1980).
4. Armstrong BG. The effects of measurement errors on relative risk regressions. Am J Epidemiol 132:1176–1184 (1990).
5. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease relationships and methods of correction. Ann Rev Publ Health 14:69–93 (1993).
6. Rosner B, Willett WS, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 8:1031–1040 (1989).
7. Huber P. The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol 1. (LeCam L, Neyman J, eds) Berkeley:University of California Press (1967).
8. Greenland S. Statistical uncertainty due to misclassification: implications for validation substudies. J Clin Epidemiol 41:1167–1174 (1988).
9. Spiegelman D, Gray R. Cost efficient study designs for binary response data with Gaussian covariate measurement error. Biometrics 47:851–869 (1991).
10. Rosner B, Willett WC. Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. Am J Epidemiol 127:377–386 (1988).
11. Peters JM. Epidemiologic investigation to identify chronic health effects of ambient air pollutants: phase I final report and phase II protocol. California Air Resources Board, Contract no. A033-186. University of Southern California, Los Angeles, July 15, 1992.